Case Study 3: fMRI Prediction

fMRI Prediction Task
Goal: Predict word stimulus from fMRI image

Classifier
(logistic regression, kNN, ...)

HAMMER
or
HOUSE

©Sham Kakade 2016
fMRI
fMRI

~1 mm resolution
~1 image per sec.
20,000 voxels/image

safe, non-invasive

measures Blood Oxygen Level Dependent (BOLD) response

Typical fMRI response to impulse of neural activity
Each stimulus repeated several times
fMRI Activation

Mean activation averaged over 60 different stimuli:

“bottle” minus mean activation:
Goal: Predict word stimulus from fMRI image

Challenges:
- $p >> N$ (feature dimension $>>$ sample size)
- Cost of fMRI recordings is high
- Only have a few training examples for each word

Classifier (logistic regression, kNN, ...)

HAMMER or HOUSE
Zero-Shot Classification

- **Goal:** Classify words not in the training set
- **Challenges:**
  - Cost of fMRI recordings is high
  - Can’t get recordings for every word in the vocabulary
Zero-Shot Classification

- **Goal:** Classify words not in the training set
- **Challenges:**
  - Cost of fMRI recordings is high
  - Can’t get recordings for every word in the vocabulary
- We don’t have many brain images, but we have a lot of info about the words and how they relate (co-occurrence, etc.)
- How do we utilize this “cheap” information?

![Classifier (logistic regression, kNN, ...)](image)

HAMMER or HOUSE
Semantic Features

Semantic feature values: "celery"
0.8368, eat
0.3461, taste
0.3153, fill
0.2430, see
0.1145, clean
0.0600, open
0.0586, smell
0.0286, touch
...
...
0.0000, drive
0.0000, wear
0.0000, lift
0.0000, break
0.0000, ride

Semantic feature values: "airplane"
0.8673, ride
0.2891, see
0.2851, say
0.1689, near
0.1228, open
0.0883, hear
0.0771, run
0.0749, lift
...
...
0.0049, smell
0.0010, wear
0.0000, taste
0.0000, rub
0.0000, manipulate
Zero-Shot Classification

- From training data, learn two mappings:
  - S: input image \(\rightarrow\) semantic features
  - L: semantic features \(\rightarrow\) word

- Can use “cheap” co-occurrence data to help learn L

![Diagram showing the process of feature extraction and classification]

Features of word \(\rightarrow\) Classifier (logistic regression, kNN, ...) \(\rightarrow\) HAMMER or HOUSE
Goal: Predict semantic features from fMRI image
Case Study 3: fMRI Prediction

Ridge, LASSO Review

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Sham Kakade
April 28, 2016
Linear Regression

- Model:

- MLE: \( \hat{\theta} = \arg \max_\theta \log p(D | \theta) \)

- Minimizing RSS = least squares regression
Ridge Regression

- Ameliorating issues with overfitting:

- New objective:

  - Solution:
Ridge Coefficient Path

• Typical approach: select \( \lambda \) using cross validation

From Kevin Murphy textbook
Case Study 3: fMRI Prediction

fMRI Prediction Results

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Sham Kakade
April 28, 2016
fMRI Prediction Results

- Palatucci et al., “Zero-Shot Learning with Semantic Output Codes”, NIPS 2009

- fMRI dataset:
  - 9 participants
  - 60 words (e.g., bear, dog, cat, truck, car, train, ...)
  - 6 scans per word
  - Preprocess by creating 1 “time-average” image per word

- Knowledge bases
  - Corpus5000 – semantic co-occurrence features with 5000 most frequent words in Google Trillion Word Corpus
  - human218 – Mechanical Turk (Amazon.com)
    - 218 semantic features (“is it manmade?”, “can you hold it?”,...)
    - Scale of 1 to 5
fMRI Prediction Results

- **First stage**: Learn mapping from images to semantic features
  - Ridge regression

- **Second stage**: 1-NN classification using knowledge base
fMRI Prediction Results

- Leave-two-out-cross-validation
  - Learn ridge coefficients using 58 fMRI images
  - Predict semantic features of 1st heldout image
  - Compare whether semantic features of 1st or 2nd heldout image are closer

Table 1: Percent accuracies for leave-two-out-cross-validation for 9 fMRI participants (labeled P1-P9). The values represent classifier percentage accuracy over 3,540 trials when discriminating between two fMRI images, both of which were omitted from the training set.

<table>
<thead>
<tr>
<th>corpus5000</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>P7</th>
<th>P8</th>
<th>P9</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>human218</td>
<td>90.3</td>
<td>82.9</td>
<td>86.6</td>
<td>71.9</td>
<td>89.5</td>
<td>75.3</td>
<td>78.0</td>
<td>77.7</td>
<td>76.2</td>
<td>80.9</td>
</tr>
</tbody>
</table>

Figure 1: Ten semantic features from the human218 knowledge base for the words bear and dog. The true encoding is shown along with the predicted encoding when fMRI images for bear and dog were left out of the training set.
fMRI Prediction Results

- Leave-one-out-cross-validation
  - Learn ridge coefficients using 59 fMRI images
  - Predict semantic features of heldout image
  - Compare against very large set of possible other words

Figure 2: The mean and median rank accuracies across nine participants for two different semantic feature sets. Both the original 60 fMRI words and a set of 940 nouns were considered.

Table 2: The top five predicted words for a novel fMRI image taken for the word in bold (all fMRI images taken from participant P1). The number in the parentheses contains the rank of the correct word selected from 941 concrete nouns in English.
Case Study 3: fMRI Prediction

LASSO Review

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Sham Kakade
April 28, 2016
Variable Selection

- Ridge regression: Penalizes large weights

- What if we want to perform “feature selection”? 
  - E.g., Which regions of the brain are important for word prediction? 
  - Can’t simply choose predictors with largest coefficients in ridge solution 
  - Computationally impossible to perform “all subsets” regression 

  - Stepwise procedures are sensitive to data perturbations and often include features with negligible improvement in fit

- Try new penalty: Penalize non-zero weights 
  - Penalty: 

    - Leads to sparse solutions 
    - Just like ridge regression, solution is indexed by a continuous param $\lambda$
LASSO Regression

- **LASSO**: least absolute shrinkage and selection operator

- New objective:
Geometric Intuition for Sparsity

Ridge Regression

Lasso
Soft Thresholding

\[
\hat{\beta}_j = \begin{cases} 
(c_j + \lambda)/a_j & c_j < -\lambda \\
0 & c_j \in [-\lambda, \lambda] \\
(c_j - \lambda)/a_j & c_j > \lambda
\end{cases}
\]

From Kevin Murphy textbook

©Sham Kakade 2016
LASSO Coefficient Path

From Kevin Murphy textbook
## LASSO Example

<table>
<thead>
<tr>
<th>Term</th>
<th>Least Squares</th>
<th>Ridge</th>
<th>Lasso</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.465</td>
<td>2.452</td>
<td>2.468</td>
</tr>
<tr>
<td>lcavol</td>
<td>0.680</td>
<td>0.420</td>
<td>0.533</td>
</tr>
<tr>
<td>lweight</td>
<td>0.263</td>
<td>0.238</td>
<td>0.169</td>
</tr>
<tr>
<td>age</td>
<td>-0.141</td>
<td>-0.046</td>
<td>0.094</td>
</tr>
<tr>
<td>lbph</td>
<td>0.210</td>
<td>0.162</td>
<td>0.002</td>
</tr>
<tr>
<td>svi</td>
<td>0.305</td>
<td>0.227</td>
<td>0.094</td>
</tr>
<tr>
<td>lcp</td>
<td>-0.288</td>
<td>-0.021</td>
<td>0.040</td>
</tr>
<tr>
<td>gleason</td>
<td>-0.212</td>
<td>0.267</td>
<td>0.133</td>
</tr>
<tr>
<td>pgg45</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Debiasing

Original (D = 4096, number of nonzeros = 160)

L1 reconstruction (K0 = 1024, lambda = 0.0516, MSE = 0.0027)

Debiased (MSE = 3.26e−005)

Minimum norm solution (MSE = 0.0292)

From Kevin Murphy textbook
Sparsistency

• Typical Statistical Consistency Analysis:
  – Holding model size \((p)\) fixed, as number of samples \((N)\) goes to infinity, estimated parameter goes to true parameter

• Here we want to examine \(p \gg N\) domains
• Let both model size \(p\) and sample size \(N\) go to infinity!
  – Hard case: \(N = k \log p\)
Sparsistency

- Rescale LASSO objective by $N$:

- Theorem (Wainwright 2008, Zhao and Yu 2006, ...):
  - Under some constraints on the design matrix $X$, if we solve the LASSO regression using

  Then for some $c_1 > 0$, the following holds with at least probability

- The LASSO problem has a unique solution with support contained within the true support
- If $\min_{j \in S(\beta^*)} |\beta_j^*| > c_2 \lambda_n$ for some $c_2 > 0$, then $S(\hat{\beta}) = S(\beta^*)$
Acknowledgements

Some material in this lecture was based on slides provided by:

- Tom Mitchell – fMRI
- Rob Tibshirani – LASSO